Exploring Bilingual Word Vectors for Hindi-English Cross-Language Information Retrieval

Vijay Kumar Sharma
Malaviya National Institute of Technology
Jaipur, India
Sharmavijaykumar55@gmail.com

Namita Mittal
Malaviya National Institute of Technology
Jaipur, India
mittalnamita@gmail.com

ABSTRACT
Todays, The internet has become a source of multi-lingual content. Users are not aware of multiple languages, so the language diversity becomes a great barrier for world communication. Cross-Language Information Retrieval (CLIR) provides a solution for this language barrier, where a user can search in his native language and get the relevant information in the required language. Currently, distributed word vector representation has a trend in various Natural Language Processing (NLP) task. These word vectors are used to identify similar contextual words. In this paper, we analyze the effectiveness of word vectors across the languages in Hindi-English CLIR. Skip-Gram Model (SGM) is used to learn bi-lingual word vectors from sentence aligned comparable corpus. IBM model is used to align the source language and target language words from sentence aligned comparable corpus. Best target language translation is selected with the help of top-n word alignments and word vectors.

CCS Concepts
• **Information systems→Information Retrieval**

Keywords
Word-Embedding; Skip-Gram Model; CLIR; Comparable Corpus; Contextual learning;

1. INTRODUCTION
Nowadays the internet has overwhelmed by multi-lingual content. The classical IR normally regards the documents and sentences in other languages as unwanted "noise" [1] . Global internet usage statistics shows that the numbers of web access by the non-English users are tremendously increased. But, all of them are not able to express their queries in English[1]. The needs for handling multiple languages introduce a new area of IR that is CLIR. CLIR provides the accessibility of relevant information in a language different than the query language [2]. In CLIR, a user query is translated by either direct translation i.e. *Dictionary-Based Translation* (DT), *Corpus-Based Translation* (CT) and *Machine Translation* (MT) or indirect translation i.e. *Latent Semantic Indexing (LSI), Explicit Semantic Analysis (ESA)* etc. [3].

ICIA-16, August 25-26, 2016, Pondicherry, India
© 2016 ACM. ISBN 978-1-4503-4756-3/16/08..$15.00
DOI: http://dx.doi.org/10.1145/2980258.2980310

[1] Internet World Stats: http://www.internetworldstats.com.

There are two types of direct translation approaches namely query translation and documents translation. A lot of computation time and space is elapsed in document translation approach so query translation approach is preferred [4]. DT approaches have issues of word translation disambiguation and dictionary coverage. MT and CT approach required a parallel corpus. Although it is very difficult to get a parallel corpus but if it is available then CT approach is very effective [5,6]. Most of the researchers were utilized parallel corpus to create a probabilistic lexicon. Giza++[2] tool is used to create a probabilistic word alignment table where each word has multiple translations associated with probability score. IBM model is used in Giza++ for word alignment [7,8]. Query words are translated based on either maximum probability score or Point-wise Mutual Information (PMI) score. Query word translation based on PMI score gives a very poor result because the probability of co-occurrence of two words at sentence level is very low. So in our implementation, we used maximum probability score. Indirect translation method like LSI uses the parallel corpus to create dual semantic space. LSI method used a relational algebra method, Singular Value Decomposition (SVD) and term-frequency matrix which is very large for the given parallel corpus, so computation cost of LSI method is very high [3].

Distributed word vector representations are exploited in various NLP task. Word vectors are learned from sentence aligned comparable corpus by using an existing Skip-Gram Model. SGM does not include the dense matrix multiplication as like in neural network architecture so SGM is efficient technique than the neural network architecture. A list of similar contextual words is learned for a word in SGM. Most of the previous work is focused on a single language where semantically similar words are identified for a given word [9, 10]. Many researchers extend this SGM from a single language to across the languages [11, 12]. A Bilingual Word Embedding Skip-Gram (BWESG) model is proposed for CLIR where dual semantic space is created by combining and shuffling the comparable sentence [23].

Our contribution in this work is to: (1) Analysis of BWESG model for Hindi-English CLIR. (2) A hybrid model is proposed which includes BWESG and IBM model of word alignment. (3) Comparison of this hybrid model with the Probabilistic Lexicon Model (PLM). Initially bilingual word vectors are learned by using BWESG model. It is assumed that the similar context words are having approximately closer word vector. So cosine similarity scores are computed between each source language word and all target language words. A maximum cosine similarity scored target language word is assigned to a source language word. This approach gives very poor results because many target language words are having the approximately similar word vectors as

[2] http://www.statmt.org/moses/giza/GIZA++.html

source language words. So a wrong target language word is assigned to a source language word due to the number of reasons which are explained here. In a sentence aligned corpus, sentence lengths across the languages are different; the number of vocabulary words and number of stop-words in source language and target language is different. So a hybrid model is proposed, where cosine similarity score is computed between a source language word and top-n target language words which are identified by using word alignment algorithm. Further, a hybrid model is compared with PLM in respect of CLIR scenario. A literature review is discussed in section 2. A hybrid model is proposed in section 3. Results and discussions are presented in section 4.

2. LITERATURE REVIEW

Jagarlamudi et al. [13] were prepared a Statistical Machine Translation (SMT) system which trained on aligned parallel sentences and a word alignment table was created. Pingali et al. [14] were used the bilingual lexicon and statistical lexicon created by parallel corpora for query translation. OOV words were transliterated by rule-based method. Mahapatra et al. [7] were used GIZA++ tool to get word alignment table from the parallel corpus and sentence word overlap score and WordNet similarity score was used for selecting the best translation. Saravanan et al. [15] were created probabilistic translation lexicon by statistical learning on parallel corpora. OOV words were handled with transliteration generation or mining technique. Surya et al. [16] and Shishtla et al. [17] were used GIZA++ tools to create word alignment table and CRF model was trained on this word alignment table for OOV word transliteration. Larkey et al. [18] were used the probabilistic dictionary for query translation. Bradford et al. [19] were used Machine Translation software to create parallel corpora and Cross-Lingual LSI method was used for CLIR. Nie et al. [20] were created probabilistic lexicon from the parallel corpus and the Probabilistic model combining with bilingual dictionary were used for query translation. Udupa et al. [21] were used GIZA++ tool to create a probabilistic lexicon and machine transliteration for OOV words.

Klementiev et al. [11] were used neural network language model for multi-task learning, where cross-lingual word vectors are induced from co-occurrence statistics in bilingual parallel data. Mikolov et al. [12] were proposed the skip-gram model for learning high quality word vectors that capture precise syntactic and semantic words relationship. Pennington et al. [10] were proposed a log-bilinear model, which involves the global matrix factorization method and local context window method. Zou et al. [22] were used a neural language model to learn bilingual word embedding and a word alignment model is used to extract the approximate translation equivalent of a source language word. Ganguly et al. [9] were proposed word embedding based language model for mono-lingual information retrieval. Vulic et al. [23] were proposed a BWESG model to learn bilingual word embeddings.

3. PROPOSED APPROACH

Source language query string is translated into the target language using a hybrid model which involves BWESG model and IBM word alignment model. Further, vector space model is used to retrieve relevant documents from target language datasets. A brief introduction of BWESG model [12, 23] and IBM model [24] is presented in sub-sections and then a hybrid model is discussed.

3.1 BWESG Model

A combined bilingual document corpus is constructed by combining comparable sentences from sentence aligned comparable corpus. A newly constructed combined bilingual document corpus $\{d_1,d_2,\ldots d_N\}$, where each document d_i contain sentences from both source language and target language documents(d_s,d_t) such as words from d_s and d_t are shuffled and the final words in d_i are $\{w_{s1},w_{t1},w_{s2},w_{t2},\ldots.w_{sn},w_{tm}\}$. The SGM is trained on the combined bilingual document corpus. The learning goal is to maximize the probability of predicting contextual words for each pivot word. the probability of predicting the context word v for a pivot word w is defined by softmax function as given in equation 1.

$$p(v \mid w) = \frac{1}{1+\exp(-\vec{w}.\vec{v})} \tag{1}$$

The BWESG model learns the word embeddings or word vectors for both source language and target language words over *dim* embedding dimension. A *dim* dimensional vector for word w is:

$$\vec{w} = [f_{w,1}, f_{w,2}, \ldots\ldots f_{w,\dim}]$$

Where $f_{w,k}$ denotes the k^{th} inter-lingual feature. Semantic similarity can be computed between the words both monolingually or across the language.

3.2 IBM Word Alignment Model

For a source language sentence s= $(s_1,s_2,\ldots.s_n)$ of length n and its translated sentence t=$(t_1,t_2,\ldots..t_m)$ of length m, translation probability for each source language word s_i to a target language word t_j with an alignment a : j → i is given in equation 2.

$$p(t,a \mid s) = \prod_{j=1}^{m} tp(t_j \mid s_{a(j)}) \tag{2}$$

Where *tp* represents the translation probability of the target language words against a source language word.

3.3 Hybrid Model for Query Translation

The hybrid model is a combination of IBM model and BWESG model. A word alignment table is created from sentence aligned comparable corpus by using an IBM model and the top-n target language words $(t_1,t_2,\ldots t_n)$ are considered as translation words for a source language word s_i. Word vectors are extracted from bilingual word vectors which are learned by SGM i.e. for each source language word s_i word vector is $\{s_{i,1},s_{i,2},\ldots s_{i,\dim}\}$ and for target language word t_j word vector is $\{t_{j,1},t_{j,2},\ldots t_{j,\dim}\}$. Best target language word among the top-n target language words is assigned to a source language word based on maximum Cosine Similarity Score (CSS) and minimum Euclidean Distance Score (EDS) as given in equation 3 and 4.

$$CSS = \frac{\sum_{k=1}^{\dim} s_{i,k}.t_{j,k}}{\sqrt{\sum_{k=1}^{\dim} s_{i,k}^2}\sqrt{\sum_{k=1}^{\dim} t_{j,k}^2}} \tag{3}$$

$$EDS = \sqrt{\sum_{k=1}^{\dim} (s_{i,k} - t_{j,k})^2} \quad (4)$$

3.4 Probabilistic Lexicon Model

Source language query string is tokenized and stopwords are eliminated to reduce noise in translation. Giza++ tool is used to create a probabilistic lexicon from the sentence aligned comparable corpus. Translation of source language query words are selected based on the maximum target language translation probability. This method is depicted in figure 1.

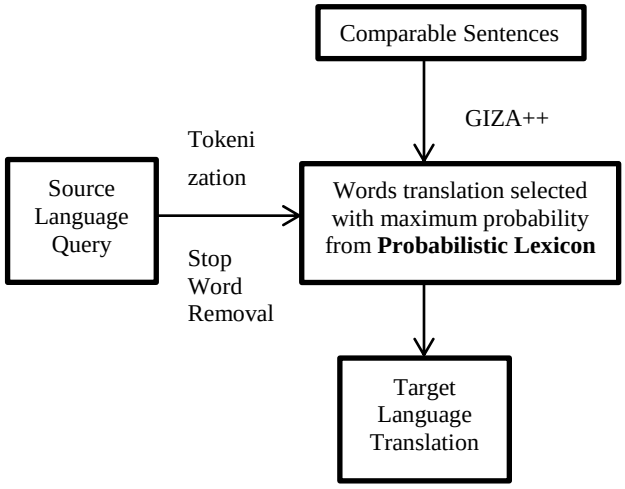

Figure. 1. PLM approach of Query Translation

4. Results and Discussions

The proposed approach is evaluated with FIRE[3] 2010 and FIRE 2011 datasets, which contains a topic set of 50 Hindi language queries and a set of target English language documents. Topic set includes <title>, <desc> and <narr> tag field in each query. We experimented with only <title> tag field. A Hindi-English parallel corpus[4] is exploited for creating word vectors and word alignments. A source language query string is translated into the target language by using the hybrid model. Vector space model is used for indexing and retrieval of target language documents. CLIR system is evaluated by using Recall and Mean Average Precision (MAP). The recall is the fraction of relevant documents that are retrieved. MAP for a set of queries is the mean of the average precision score of each query. Precision is the fraction of retrieved documents that are relevant to the query. Comparative result analysis of the proposed hybrid model and the PLM is presented in table 1. Top-5 and top-10 target language words are selected for a source language word by using IBM model. CSS and EDS measures are used for selecting best translation among top-5 and top-10 translations.

BWESG without using top-n word alignment gives a very poor result as explained in the introduction section. So a hybrid model is proposed which includes IBM word alignment model and top-5 and top-10 words are tested to identify the best translation. CSS and EDS measurements are calculated between source language word vector and top-n target language word vectors. Hybrid model with top-5 gives the better MAP than the top-10 because numbers of target language words are reduced. EDS measure gives the better MAP than the CSS measure. So finally, Hybrid model gives the better MAP with top-5 word and EDS measure as reported in table 1. Although PLM approach achieves better MAP than the hybrid model, but our objective is to exploring the word vectors representation across the languages.

Table 1. Comparative analysis of Hybrid model and PLM

Experiment	Fire 2010		Fire 2011	
	Recall	**MAP**	**Recall**	**MAP**
Hybrid Model+ Top-5+CSS	0.5727	0.1223	0.4643	0.1053
Hybrid Model+ Top-10+CSS	0.4747	0.1059	0.4060	0.0758
Hybrid Model+ Top-5+EDS	0.6034	0.1382	0.5418	0.1085
Hybrid Model+ Top-10+EDS	0.5329	0.1144	0.4889	0.0866
PLM	0.7488	0.2267	0.6791	0.1672

5. Conclusion

Word vectors learning for identifying similar contextual words perform well for a single language but for the cross-language scenario, it will generate non-contextual words. The hybrid model reduces the probability of testing the target language word vectors by selecting top-n words. So Hybrid model with top-5 and EDS measure, perform better than the basic BWESG model. PLM model selects the maximum probability target language translation from the probabilistic lexicon. Although it is very straightforward from the experimental results analysis that the PLM approach performs better than the hybrid model but our objective is not to show the superiority of the PLM approach over the word vectors learning. Our objective is to analysis and renovation of word vectors learning with respect to Hindi-English CLIR. So in future, we will improve the word vectors learning and improve the proposed hybrid model.

6. REFERENCES

[1] Mustafa A, Tait J, and Oakes M. Literature review of cross-language information retrieval. In *Transactions on Engineering, Computing and Technology, ISSN*. (2005).

[2] Nagarathinam A, and Saraswathi S. State of art: Cross Lingual Information Retrieval System for Indian Languages. In *International Journal of computer application*. Vol. 35, No. 13, (2011) , 15-21.

[3] Wang A, Li Y, and Wang W. Cross language information retrieval based on lda. In *International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009. IEEE*, vol. 3, 485-490.

[4] Nasharuddin NA, and Abdullah MT. Cross-lingual Information Retrieval State-of-the-Art. In *electronic Journal of Computer Science and Information Technology (EJCSIT)*. Vol. 2, No. 1, (2010), 1-5.

[5] Sharma VK, and Mittal N. Cross Lingual Information Retrieval (CLIR): Review of Tools, Challenges and

[3] http://fire.irsi.res.in/fire/home

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0

Translation Approaches. In *Information System Design and Intelligent Application,* (2016), 699-708.

[6] Sujatha P, and Dhavachelvan P. A review on the Cross and Multilingual Information Retrieval. In *International Journal of Web & Semantic Technology (IJWesT)*.Vol.2, No.4, (2011), 155-124.

[7] Mahapatra L, Mohan M, Khapra MM, Bhattacharyya P. OWNS: Cross-lingual word sense disambiguation using weighted overlap counts and wordnet based similarity measures. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, (2010), 138-141.

[8] Shishtla P, Surya G, Sethuramalingam S, Varma V. A language-independent transliteration schema using character aligned models at NEWS 2009. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Association for Computational Linguistics, (2009), 40-43.

[9] Ganguly D, Roy D, Mitra M, and Jones G. A word embedding based generalized language model for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, (2015), 795-798.

[10] Pennington, J., Socher, R., and Manning, C. D. Glove: Global Vectors for Word Representation. In *EMNLP*, Vol. 14, (2014, October), 1532-1543.

[11] Klementiev, A., Titov, I., and Bhattarai, B. Inducing crosslingual distributed representations of words. In *Saarland Univerisity, Germany,* (2012).

[12] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, (2013), 3111-3119.

[13] Jagarlamudi J, and Kumaran A. Cross-Lingual Information Retrieval System for Indian Languages. In *Advances in Multilingual and Multimodal Information Retrieval*, Springer Berlin Heidelberg, (2007), 80-87.

[14] Pingali P, Jagarlamudi J, and Varma V. A Dictionary Based Approach with Query Expansion to Cross Language Query Based Multi-Document Summarization: Experiments in Telugu-English. *Mumbai, India,* (2008).

[15] Saravanan K, Udupa R, and Kumaran A. Crosslingual information retrieval system enhanced with transliteration generation and mining. In *Forum for Information Retrieval Evaluation (FIRE-2010) Workshop* (2010).

[16] Surya G, Harsha S, Pingali P, and Verma V. Statistical transliteration for cross language information retrieval using HMM alignment model and CRF. In *Proceedings of the 2nd Workshop on Cross Lingual Information Access* (2008).

[17] Shishtla P, Surya G, Sethuramalingam S, and Varma V. A language-independent transliteration schema using character aligned models at NEWS 2009. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Association for Computational Linguistics, (2009), 40-43.

[18] Larkey LS, Connell ME, Abduljaleel N. Hindi CLIR in thirty days. *ACM Transactions on Asian Language Information Processing (TALIP),* Vol. 2, no. 2, (2003), 130-142.

[19] Bradford R, and Pozniak J. Combining Modern Machine Translation Software with LSI for Cross-Lingual Information Processing. In *2014 11th International Conference on Information Technology: New Generations (ITNG)*, IEEE, (2014), 65-72.

[20] Nie J, Simard M, Isabelle P, and Durand R. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, (1999), 74-81.

[21] Udupa R, Jagarlamudi J, and Saravanan K. Microsoft research india at fire2008: Hindi-english cross-language information retrieval. In *Working notes for Forum for Information Retrieval Evaluation (FIRE) Workshop* (2008).

[22] Zou, W. Y., Socher, R., Cer, D. M., and Manning, C. D. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP* (2013), 1393-1398.

[23] Vulic I., and Moens M. F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In*Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM (2015, August), 363-372.

[24] Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. Vol. 999. *Cambridge: MIT press*, (1999).